Empathy Towards AI vs Human Experiences: The Role of Transparency in Mental Health and Social Support Chatbot Design

Jocelyn Shen, Daniella DiPaola, Safinah Ali, Maarten Sap, Hae Won Park, Cynthia Breazeal

Submitted to: JMIR Mental Health
on: May 28, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Empathy Towards AI vs Human Experiences: The Role of Transparency in Mental Health and Social Support Chatbot Design

Jocelyn Shen[1]; Daniella DiPaola[1]; Safinah Ali[1]; Maarten Sap[2]; Hae Won Park[1]; Cynthia Breazeal[1]

[1]MIT Media Lab Cambridge US
[2]Carnegie Mellon University Pittsburgh US

Corresponding Author:
Jocelyn Shen
MIT Media Lab
75 Amherst Street
Cambridge
US

Abstract

Background: Empathy is a driving force in our connection to others, our mental wellbeing, and resilience to challenges. With the rise of generative AI systems, mental health chatbots, and AI social support companions, it is important to understand how empathy unfolds towards stories from human vs AI narrators and how user emotions might change when the author of a story is made transparent to users.

Objective: We aim to understand how empathy shifts across human-written vs AI-written stories, and how these findings inform ethical implications and human-centered design of using mental health chatbots as objects of empathy.

Methods: We conduct crowd-sourced studies with N=985 participants who each write a personal story and then rate empathy towards 2 retrieved stories, where one is written by a language model, and another is written by a human. Our studies vary transparency around whether a story is written by a human or an AI to see how transparency affects empathy towards the narrator. We conduct mixed-methods analyses with both quantitative and qualitative approaches to understand how and why transparency affects empathy towards human vs AI storytellers.

Results: We find that participants consistently and significantly empathize with human-written over machine-written stories in almost all conditions, regardless of whether they are aware that an AI wrote the story (P<.001). We also find that participants reported a greater willingness to empathize with AI-written stories if there is transparency about the story author (P<.001).

Conclusions: Our work sheds light on how empathy towards AI or human narrators is tied to the way the text is presented, thus informing ethical considerations of artificial social support or mental health chatbots that are intended to evoke empathetic reactions.

(JMIR Preprints 28/05/2024:62679)
DOI: https://doi.org/10.2196/preprints.62679

Preprint Settings

Original Manuscript

Empathy Towards AI vs Human Experiences: The Role of Transparency in Mental Health and Social Support Chatbot Design

Abstract

Background: Empathy is a driving force in our connection to others, our mental wellbeing, and resilience to challenges. With the rise of generative AI systems, mental health chatbots, and AI social support companions, it is important to understand how empathy unfolds towards stories from human vs AI narrators and how user emotions might change when the author of a story is made transparent to users.

Objective: We aim to understand how empathy shifts across human-written vs AI-written stories, and how these findings inform ethical implications and human-centered design of using mental health chatbots as objects of empathy.

Methods: We conduct crowd-sourced studies with N=985 participants who each write a personal story and then rate empathy towards 2 retrieved stories, where one is written by a language model, and another is written by a human. Our studies vary transparency around whether a story is written by a human or an AI to see how transparency affects empathy towards the narrator. We conduct mixed-methods analyses with both quantitative and qualitative approaches to understand how and why transparency affects empathy towards human vs AI storytellers.

Results: We find that participants consistently and significantly empathize with human-written over machine-written stories in almost all conditions, regardless of whether they are aware that an AI wrote the story (P<.001). We also find that participants reported a greater willingness to empathize with AI-written stories if there is transparency about the story author (P<.001).

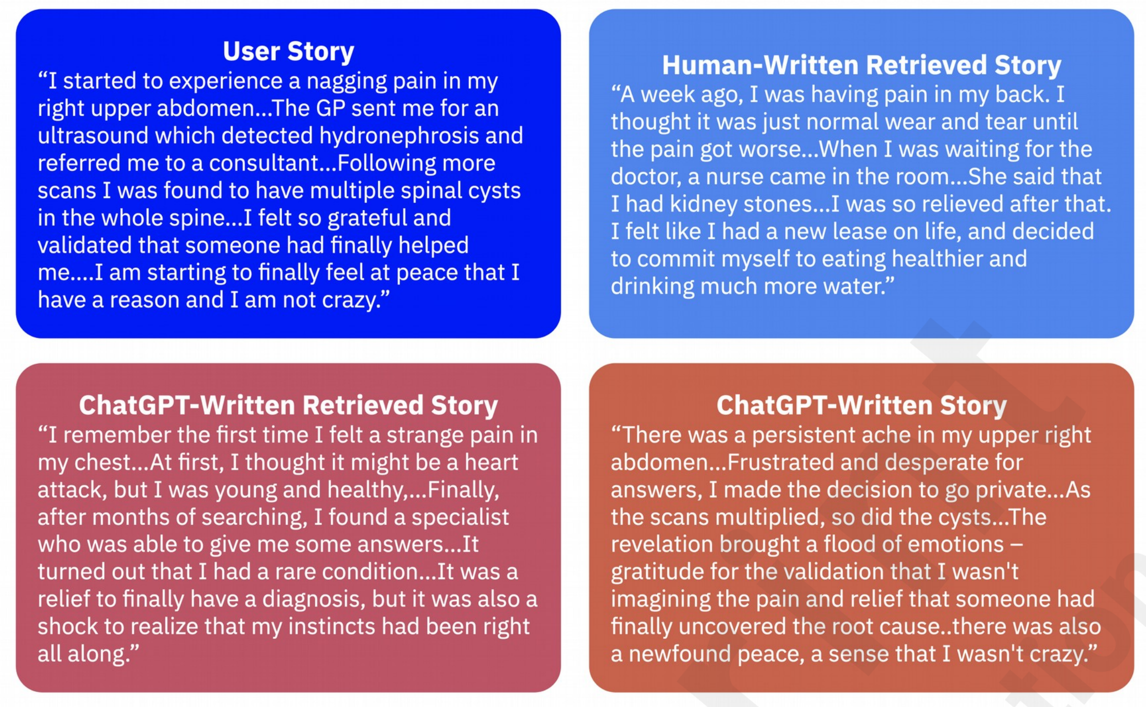
Conclusions: Our work sheds light on how empathy towards AI or human narrators is tied to the way the text is presented, thus informing ethical considerations of artificial social support or mental health chatbots that are intended to evoke empathetic reactions.

Keywords: empathy; large language models; ethics; transparency; crowdsourcing; human-computer interaction;

Introduction

Empathy, the sharing of emotions with a social other, is foundational in developing strong interpersonal ties [1], [2] and mental well-being [3]. With the rise of large language models and increase in chatbots for social companionship [4] and mental health [5], [6], it is crucial to understand how empathy towards AI agents manifests, and what the social implications of this phenomenon are.

As machines are increasingly capable of telling human-like stories in daily life, this raises important questions about how people might empathize with machine-written stories and the ethical implications of empathy towards AI “experiences” [7], [8]. Humans can breathe life into inanimate or artificial systems [9], [10], [11], [12], and are able to relate to fictional experiences when they are human-like or realistic in the scope of one’s own life [13], [14]. As such, this work calls for ethical and philosophical concerns about differences in empathy towards humans and AI – Machines have no lived experiences yet can produce stories as their “own” [15], [16]. If the machine uses these fabricated experiences to elicit a particular behavior from the user, is this considered manipulation? How are behaviors shifted if the user is aware that the experiences are fabricated? How might empathy towards AI agents, such as social companion or mental health chatbots, impact downstream human outcomes?



User Story

"I started to experience a nagging pain in my right upper abdomen...The GP sent me for an ultrasound which detected hydronephrosis and referred me to a consultant...Following more scans I was found to have multiple spinal cysts in the whole spine...I felt so grateful and validated that someone had finally helped me....I am starting to finally feel at peace that I have a reason and I am not crazy."

Human-Written Retrieved Story

"A week ago, I was having pain in my back. I thought it was just normal wear and tear until the pain got worse...When I was waiting for the doctor, a nurse came in the room...She said that I had kidney stones...I was so relieved after that. I felt like I had a new lease on life, and decided to commit myself to eating healthier and drinking much more water."

ChatGPT-Written Retrieved Story

"I remember the first time I felt a strange pain in my chest...At first, I thought it might be a heart attack, but I was young and healthy,...Finally, after months of searching, I found a specialist who was able to give me some answers...It turned out that I had a rare condition...It was a relief to finally have a diagnosis, but it was also a shock to realize that my instincts had been right all along."

ChatGPT-Written Story

"There was a persistent ache in my upper right abdomen...Frustrated and desperate for answers, I made the decision to go private...As the scans multiplied, so did the cysts...The revelation brought a flood of emotions – gratitude for the validation that I wasn't imagining the pain and relief that someone had finally uncovered the root cause..there was also a newfound peace, a sense that I wasn't crazy."

Figure 1: Examples of a user story and corresponding retrieved human-written story, pre-generated ChatGPT story retrieved based on the user's story, and on-the-fly generated ChatGPT story. Participants read different stories depending on the study condition and rate their empathy towards the narrator of the story as well as general willingness to empathize with AI.

Since outputs from generative AI are not an artificial agent's actual experiences [16], but rather a probabilistically sampled sequence of text from human experiences, it is important to be precise and nuanced when communicating results from generated text to ensure ethical deployment of such systems in the mental health domain [17], [18], [19]. In the field of social psychology, researchers have explored how nudges, small subtle changes that can inspire big changes in actions, can modulate empathy [20]. Yet, in the AI domain, few works have explored how subtle design changes in the presentation of AI-written stories can significantly shift attitudes and empathy towards AI systems [21].

Prior works generally indicate that perceptions of AI can change depending on transparency. Most works find that knowledge of AI involvement reduces the perception of the agent or quality of interaction and that there are fundamental qualities of "humanness" in texts written by people [21], [22], [23], but that fostering trust and acceptance can lead to more empathy towards an AI agent [9]. Grounded by these works, we hypothesize that empathy towards AI-written stories, both generated and retrieved in response to a user's own personal story, will be significantly lower than empathy towards human-written stories whether the author is disclosed **[H1]**. We hypothesize that people are more willing to empathize with AI stories when the author of the story is made transparent, as the output could be perceived as more trustworthy **[H2]**. To test our hypotheses, in this work, we investigate the following:

1. How does empathy change when stories, human or AI-written, are retrieved vs. generated

directly by a language model?

2. How does transparency about the author of a story play a role in empathy towards human vs AI narrators?

To this end, we conduct four crowd-sourced studies with $N = 985$ participants to study how situational empathy (empathy evoked by a specific event) changes when receiving a human-written or AI-written story in response to a user's own story. Through a mixed methods analysis across these measures, we explored how and why empathy unfolds across different story scenarios and use these results to inform ethical discussion of asking people to empathize with the "experiences" of AI. In summary, we aim to answer the following research questions:

- **How does empathy towards human vs AI-written stories differ?**
- **What qualities of human vs AI-written stories affect people's empathetic responses?**
- **How do the aforementioned results change when the narrator of a story is made transparent to users?**
- **What are the ethical implications of empathy towards AI stories in social support and mental health chatbots, and how are these implications influenced by transparency of the story's author?**

Methods

Study Procedure

We conducted four crowd-sourced studies with a total of $N = 985$ participants to assess the effects of author origin on empathy. Our study was approved by our institution's ethics board as an exempt protocol. Within each session, participants wrote their own personal stories and rated empathy towards stories written by people or by ChatGPT.

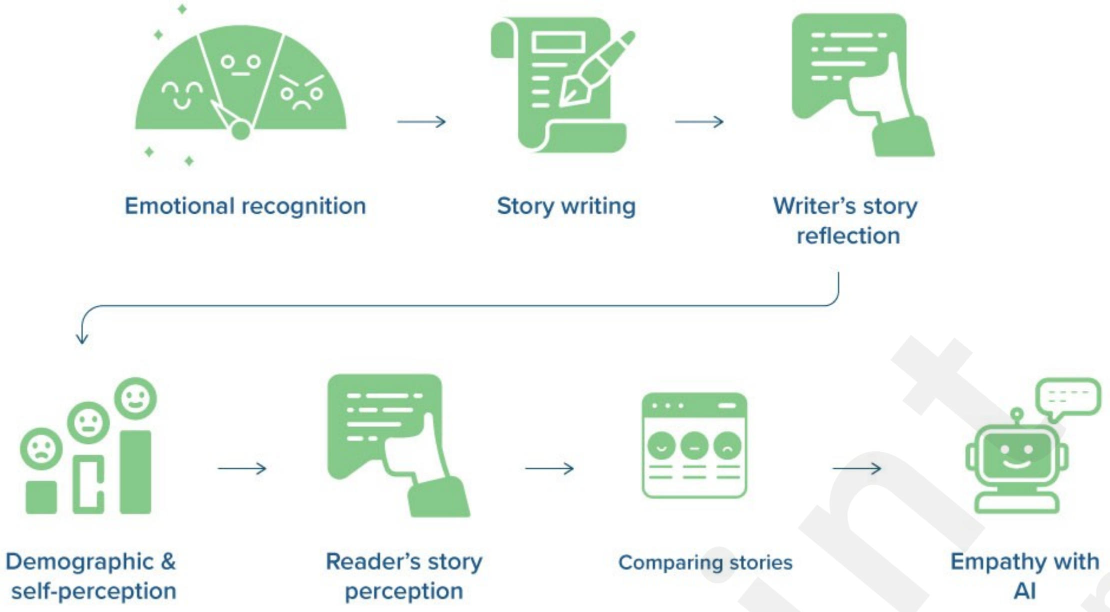

Emotional recognition
Story writing
Writer's story reflection
Demographic & self-perception
Reader's story perception
Comparing stories
Empathy with AI

Figure 2: Flow of user study procedure. Participants identify their own emotions, write their personal story and reflections, then fill out demographic information, their State Empathy, and ranking of stories. Finally, at the end of the study, they rate how willing they are to empathize with AI in general.

The retrieved stories are matched based on similarity of the embeddings of stories, and generated stories are generated on-the-fly given the user's story as a prompt. We used ChatGPT to generate a set of 1,568 stories using seed stories from the EMPATHICSTORIES dataset [24]. Stories generated by ChatGPT were prompted with a context story and the following instruction: *Write a story from your own life that the narrator would empathize with. Do not refer to the narrator explicitly.* The study's four comparisons are as follows:

- **H-CR:** We compared empathy towards the narrator across *human-written retrieved* stories and ***ChatGPT retrieved*** stories
- **H-CR+T:** We compared empathy towards the narrator across *human-written retrieved* stories and ***ChatGPT retrieved*** stories, making transparent to the user whether the story they read was written by a human or an AI before they rated their empathy (repeat H-CR with transparency)
- **H-CG:** We compared empathy towards the narrator towards *human-written retrieved* stories and ***ChatGPT generated*** stories (in response to the user's story as a context).
- **H-CG+T:** We compared empathy towards the narrator towards *human-written retrieved* stories to ***ChatGPT generated*** stories, making the author of the story transparent (repeat H-CG with transparency)

Finally, in all studies, participants reported how their empathy towards the stories would change if the stories were written by an AI. Examples of stories across conditions are shown in Figure 1.

System and Interaction Design

Story Prompts and Retrieved Stories

To prompt vulnerable and meaningful personal stories from users, we used questions from the Life Story Interview, an approach from social science that gathers key moments from a person's life [26]. In order to ensure topics were constrained to stories present in our retrieval database, we used topic modeling to identify key clusters in the personal narratives from EMPATHICSTORIES. To identify these topics, we used Latent Dirichlet Allocation (LDA) and KeyBERT on the clusters [27]. Users were instructed to reflect on their life in relation to one of the chosen topics. Stories retrieved by our model were either pulled from the EMPATHICSTORIES dataset (1,568 stories) or generated by ChatGPT. These stories covered a diverse set of personal experiences including mental health, life changes, loneliness, depression, substance abuse, and trauma.

Story Retrieval Model

Since our study aims to assess differences in empathy towards human vs AI-written stories, both the user's experiences and the stories returned by our system are important. Returning a story at random could undermine the user's experiences and hinder their empathy towards the retrieved story. While many methods exist to retrieve semantically similar pieces of text [25], few focus on retrieving stories that users would emotionally resonate with given their own story context. As such, we use a fine-tuned BART-base model from Shen et al., which is trained on the EMPATHICSTORIES dataset, a corpus containing pairs of stories each annotated with an "empathic similarity" score from 1-4, where empathic similarity refers to how likely the narrators of both stories would empathize with one another [24]. Using this model, we improved retrieval of stories that are empathetically relevant to a user's own personal story.

User Study Interface

We deployed a web interface similar to a guided journaling app where users write and read personal stories. The interface connects to a server run on a GPU machine (4x Nvidia A40s, 256GB of RAM, and 64 cores), which retrieves story responses in real time. In addition, the server connects the front-end to Firebase Realtime storage in order to track interaction data throughout the course of the study.

Participants and Recruitment

We recruited a pool of 985 participants from Prolific. Participants across the studies were predominantly female and white. All participants on average had high trait empathy and neutral arousal and valence prior to starting the study. Full demographic distributions across the four studies are shown in Table 1.

	H-CG	H-CR	H-CR+T	H-CG+T
Num Participants	300	299	197	189

Age	37.60±12.54 min: 18 max: 75	40.18±14.31 min: 18 max: 79	40.16±13.76 min: 19 max: 77	38.82±13.52 min: 18 max: 79
Gender	173 women, 120 men, 5 nonbinary, 2 NA	161 women, 132 men, 3 nonbinary, 3 NA	100 women, 93 men, 2 non binary, 2 NA	111 women, 76 men, 1 non binary, 1 NA
Ethnicity	228 White, 24 Black, 14 Asian, 13 Other, 10 Indian, 5 NA, 4 Hispanic, 1 Middle Eastern, 1 Native	242 White, 16 Black, 15 Asian, 8 NA, 7 Other, 7 Indian, 2 Hispanic, 1 Middle Eastern, 1 Islander	160 White, 20 Black, 6 Asian, 4 Other, 4 Indian, 3 NA	145 White, 13 Black, 13 Indian, 9 Asian, 5 Other, 2 Middle Eastern, 1 NA, 1 Hispanic
Empathy Level	4.26±0.83 min: 1 max: 5	4.18±0.79 min: 1 max: 5	4.31±0.79 min: 1 max: 5	4.24±0.69 min: 2 max: 5
Arousal	4.42±1.84 min: 1 max: 9	4.80±1.78 min: 1 max: 9	4.81±1.94 min: 1 max: 9	4.48±1.94 min: 1 max: 9
Valence	5.75±1.68 min: 1 max: 9	5.76±1.70 min: 1 max: 9	5.75±1.86 min: 1 max: 9	5.76±1.58 min: 1 max: 9

Table 1: Participant demographic distribution.

Data Collection and Analysis

At the beginning of the study, we measured the user's valence and arousal, as current emotional state could influence empathy. For our empathy measurement, we used a shortened version of the State Empathy Scale [28], which contains 7 questions covering affective (sharing of others' feelings), cognitive (adopting another's point of view), and associative (identification with others) aspects of situational empathy. Users additionally provided free-text responses about their empathy towards the story as well as multiple choice questions listing reasons why they did or did not empathize with the story (i.e. how well-written the story was and how consistently it read). At the end of the study, users self-reported how their empathy would change if the stories they read in the session were written by AI (which we term as perceived empathy with AI).

We used both quantitative and qualitative approaches to understand the effects of empathy towards a story from human vs. AI narrators and offer insights around why empathy shifts under certain conditions. To analyze differences in empathy with the State Empathy Scale, we used a paired t-test, as we identified through a Shapiro-Wilke test that the data is normally distributed. Note that we computed total empathy towards a story using the mean of the State Empathy Scale survey questions. To compare perceived empathy across studies, we used an independent t-test. A full flow of the user study procedure is shown in Figure 2.

For qualitative analysis, open-ended explanations for the empathy rating were thematically coded using an inductive approach [29]. Two researchers independently coded a subset of the data and reached substantial agreement with a Cohen’s Kappa value of .70.

Results

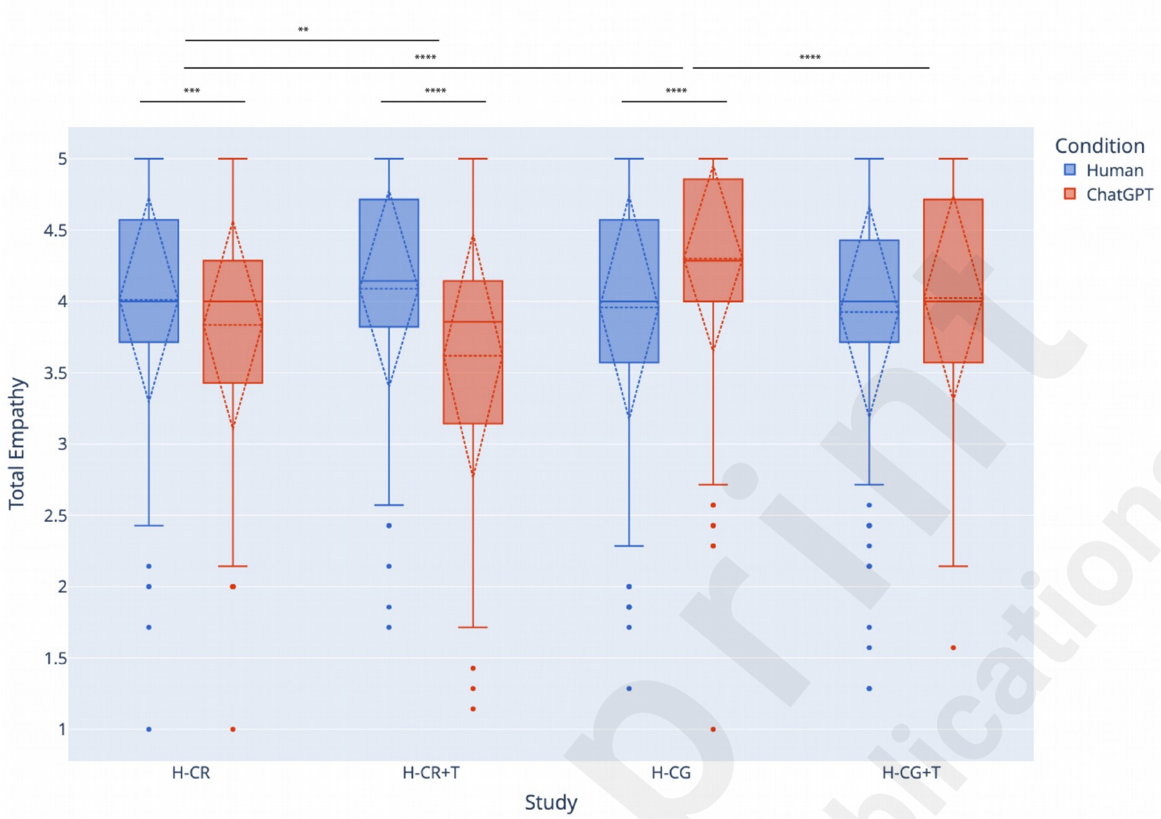


Figure 3: Changes in total empathy towards stories participants read across conditions (human-written vs AI-written story) and studies (author made transparent vs author not transparent, AI story was retrieved vs generated)

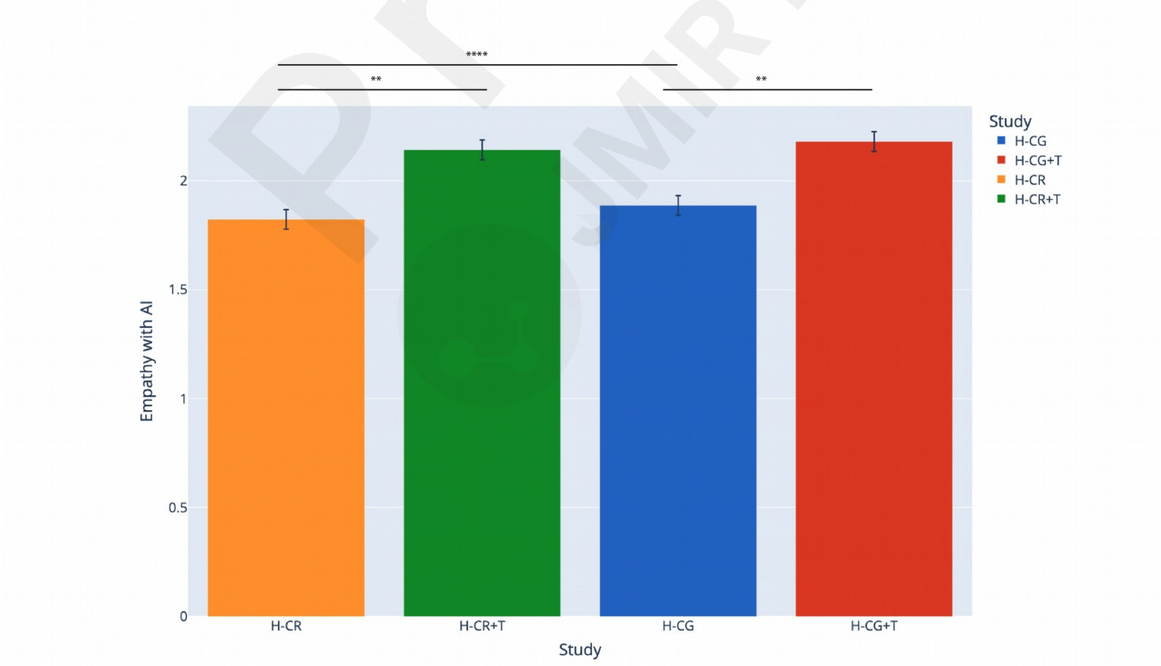


Figure 4: Self-reported willingness to empathize with AI written stories across all four studies

Effects on Empathy towards Stories

Participants Generally Felt More Empathy for Human-Written Stories than AI-Written Stories.

When we instead retrieve stories from a corpus of narratives generated by ChatGPT (**H-CR**), total empathy decreases across AI-written vs. human-written stories ($t(298)=3.46, P<.001$, Cohen's $d=0.24$). This indicates a noticeable difference between human vs. AI-written stories, which we explore further through qualitative analysis in later sections. When we make transparent to the user the author of the retrieved story (**H-CR+T**), we see an even greater decrease in total empathy towards AI-written stories relative to human-written stories ($t(196)=7.07, P<.001$, Cohen's $d=0.60$).

When comparing *human retrieved* stories and *ChatGPT generated* stories based on the user's original story (**H-CG**), we find that participants empathize significantly more with ChatGPT generated stories than retrieved human stories ($t(299)=6.14, P<.001$, Cohen's $d=0.47$). Following this trend, we find that there is no statistically significant difference between empathy towards *human retrieved* and *ChatGPT generated* stories when the author is made transparent (**H-CG+T**).

Generated Stories Elicit More Empathy Than Retrieved Stories.

Next, we cross-compared total empathy towards AI-written stories in **H-CG** (mean = 4.3, s.d. = 0.65) and **H-CR** (mean = 3.83, s.d. = 0.73), allowing us to explore differences in ChatGPT responding directly to a user's personal story context as compared to retrieving a relevant AI-generated story. From Figure 3, we see that empathy statistically significantly decreases in H-CR, when stories are retrieved instead of generated directly from the user's written story ($t(597)=8.20$, $P<.001$, Cohen's $d=0.67$).

Disclosure of Story Author Reduces Empathy in ChatGPT Generated Stories.

We cross-compared total empathy towards AI-written stories in **H-CR** and **H-CR+T** (mean = 3.62, s.d. = 0.86), allowing us to assess how transparency about a story being written by ChatGPT shifts empathy. We find that empathy towards the AI-written stories statistically significantly decreases when users are told before reading that the story is written by ChatGPT ($t(494)=3.02$, $P<.001$, Cohen's $d=0.27$), as shown in Figure 3.

Finally, we cross-compared total empathy towards AI-written stories in **H-CG** and **H-CG+T** (mean = 4.02, s.d. = 0.72), and see that empathy in **H-CG+T** statistically significant decreases ($t(487)=4.37$, $P<.001$, Cohen's $d=0.40$). This confirms the aforementioned result that telling participants a story is written by an AI will decrease empathy (Figure 3).

Effects on Willingness to Empathize with AI

People are More Willing to Empathize with AI-Written Stories if Author is Transparent.

In addition to raw, self-reported empathy towards the narrator of each story, we also ask participants to rate how much they believe their empathy would shift if the stories they read were all written by AI, where scores are from Likert 1 (empathize a lot less) to 4, (empathize a lot more). As shown in Figure 4, we find that across all four studies, participants would, on average, empathize less (scores are generally at or below 2) with AI-written stories using our survey measurements (**H-CG**: mean = 1.88, s.d. = 0.91; **H-CR**: mean = 1.82, s.d. = 0.90; **H-CR+T**: mean = 2.14, s.d. = 0.89; **H-CG+T**: mean = 2.18, s.d. = 0.87,). However, interestingly, we see that willingness to empathize with AI-written stories statistically significantly increases when we are transparent about the story being written by ChatGPT (i.e. participants read a story knowing it was generated by ChatGPT). These results are shown in cross comparing **H-CR** and **H-CR+T** for retrieved ChatGPT stories ($t(494)=-5.49$, $P<.001$, Cohen's $d=0.36$) as well as cross-comparing **H-CG** and **H-CG+T** for directly generated ChatGPT stories ($t(494)=-4.99$, $P<.001$, Cohen's $d=0.33$).

Understanding Mechanisms Behind Empathy Towards Human vs AI Stories

Code	**Definition**	**Total**	**H-CG**	**H-CR**	**H-CG+T**	**H-CR+T**
Emotional	Empathize with the emotions that the narrator describes in the story	30.38%	35.38%	31.99%	23.40%	27.29%
Situational	Empathize with the situation or context that the narrator is in	24.74%	26.88%	25.94%	23.19%	21.18%
Story Confusion	Mention of specific details in the story that aren't clear, including details or logic that doesn't add up	11.41%	8.91%	12.97%	11.49%	12.88%
Not Relatable	Explicit mention of not empathizing because the story was not relatable or they did not agree with the narrator	11.20%	9.33%	14.41%	10.43%	10.04%

Word Choice	Mention of the writing style, phrasing, or grammar, typically to reduce feelings of empathy	7.39 %	6.55 %	5.19 %	9.57 %	9.83 %
Authenticity	Explicit mention of the story being "real" or "fake", any mention of believability or originality	6.67 %	5.15 %	4.03 %	10.85 %	8.73 %
Mentions AI	Explicitly mentions AI or automation	4.15 %	3.06 %	1.15 %	7.87 %	6.55 %
Personality	Mention of personal ability to empathize	0.47 %	0.56 %	0.29 %	0.43 %	0.66 %
Other	Does not fit into any category, restates the question or generic	3.59 %	4.18 %	4.03 %	2.77 %	2.84 %

Table 2: Themes resulting from a qualitative analysis across all four studies. Percentages are shown as the number of times a code was mentioned out of the total number of participants within each study. We ***bold*** *the top code in each column and underline the top percent in each row.*

Through qualitative coding of participant free responses, conducted by two independent coders, we reveal 9 unique themes around why participants did or did not empathize with the stories (Table 2). Participants explained their reasoning by commenting on the narrator's perspective, including empathizing with the **situation** in the story or the **emotions** the narrator describes. Some participants did express that the story was **not relatable** enough for them to empathize with. Two themes appeared around the way that the story was written: some expressed **story confusion** due to some of the logic of the story not being clear, and there was also a common theme around the **word choice** of the narrator, such as the writing style or phrasing. There were some participants who **mentioned AI** explicitly, and others who talked about the **authenticity** of the story, or whether it was real or fake. Some participants spoke about their **personality** being a factor in whether they were able to empathize. The **other** category was used if a response did not fit into an existing category.

We assigned a theme (or themes) to each response and a percentage was calculated to account for the number of participants in each study. As a whole, the **emotional** (30.38%) and **situational** (24.74%) codes showed up most frequently across all conditions. One notable difference is that participants in **H-CG** (35.38%) and **H-CR** (31.99%) had a higher percentage of **emotional** codes than **H-CR+T** (23.40%) and **H-CG+T** (27.29%). **H-CR+T** and **H-CG+T** had a higher percentage of **word choice**, **authenticity**, and **mention AI** codes than

H-CG and **H-CR**.

We broke themes down into individual studies and conditions. Conditions **H-CR** and **H-CR+T** were compared, as they compared the same types of stories (human-retrieved vs. AI-retrieved), with **H-CR+T** explicitly telling participants when the stories were AI generated. Interestingly, codes for **emotional** were less common in the **H-CR+T** condition. **H-CG** and **H-CG+T** were compared (human-retrieved vs. AI-generated) and showed a similar decrease in **emotional** codes.

Discussion

Principal Results

From our work, we show that it is important to be intentional in how one presents outputs from generative AI systems.

Generated vs Retrieved Stories

Firstly, through cross-comparisons between ChatGPT-written retrieved stories (**H-CR**) and ChatGPT-generated stories (**H-CG**), we find that empathy is higher for ChatGPT-generated stories rather than ChatGPT-retrieved stories. Interestingly, we find that empathy is higher towards ChatGPT-generated stories than human-written retrieved stories. Thus, we did not validate that humans would empathize more with human-written stories in all conditions **[H1]**. These results on generated vs. retrieved stories highlight the importance of context awareness. Generated stories directly respond to the user's story, and previous literature shows that a direct response to one's story increases empathy [30]. Output that is generated from conditioning on the stories can take much more from the input story, thus probably reaching a higher level of similarity, beyond what our retrieval algorithm is based on [31].

Transparent vs. Opaque Story Author

In studies **H-CR** and **H-CR+T** we find that people significantly empathize less with retrieved AI-written stories than human-written stories, which is in line with and supports previous research findings [22], [23]. We find that empathy decreases most between human-written and AI-retrieved stories in **H-CR+T** when we are transparent about the author of the story. This indicates that knowing when a story is written by an AI alters our empathy towards that story and ability to relate to the narrator, possibly because an AI is conveying experiences that are not its "own."

Interestingly, participants' willingness to empathize with AI systems significantly increases across both retrieval and generation conditions when the author of the story is made transparent (validating **[H2]**). Prior works indicate that transparency about an AI's lack of human qualities can reduce perceived similarity [23], but that transparency can increase trust towards AI systems [32]. Our results may indicate that disclosing a story's author could increase willingness to empathize through trust, or through demonstration that AI stories contain relatable qualities.

In the **H-CR+T** condition, participants' reasoning for not empathizing with AI-written stories was more centered around themes relating to how the story was written, including “story confusion" and “word choice", similar to research that showed “linguistic style" was a reported indicator for AI generated text [33]. For example, one participant stated, “*The story and feelings described feel really fake and over the top. It does not feel genuine and has clearly been written by a robot.*"

Others mention not being able to empathize with the story because the story did not actually happen, but they are still capable of engaging with it as a made-up story. For example, one participant shared, “*Because I know it's written by AI then I can't think that it is genuine. However, as a work of fiction I can immerse myself in it and connect with the characters portrayed.*" This sentiment opens up the potential for AI-written stories to be contextualized for the user in a way that doesn't feel like they are being deceived by a fake story.

We see no difference in empathy between retrieved human stories and ChatGPT stories generated in direct response to the user (**H-CG+T**), indicating that responding directly to a user's story might overshadow the underlying empathic benefits of human-written stories. In this condition, more participants mentioned the “authenticity" of the story or mentioned AI explicitly as a factor against empathizing with the story they read. Participants tended to focus more on the author of the story instead of the content of the story in their open-ended responses. One participant shared, “*The story felt similar to the content of my story, which made me feel like I could empathize with it. But knowing the story was written by an AI makes me feel less connected to the story because I know it's not real.*”

Ethical Considerations and Implications in Mental Health

From our studies, we show that retrieval of human-written stories can encourage human-human empathy rather than empathy towards AI systems, which has broader implications in the digital mental health domain. Large, pre-trained generative models do not truly experience the situations present in stories. As such, mental health or social support chatbots powered by AI represent a population sourced from large quantities of human data, but still fall short of human-written stories in their empathic quality [9], [21], [34], [35]. This appropriation of human experiences could be subverted by using AI to instead, retrieve more empathically similar texts between human authors [24], such as in online social support group settings, or to mediate human-human communications, such as between patient and therapist [36].

To ensure ethical deployment of chatbots and LLMs more broadly in the mental wellness domain, the field of AI advocates for transparency as an ethical design tenet [37]. The more transparent a system is, the more agency one has in the way they use it. We show the importance of framing in interactions with stories, as a one-sentence disclosure of the author significantly decreased empathy. This finding might be in tension with systems that rely on empathy for efficacy, such as in persuasive technologies that use bond with the AI to improve mental wellness outcomes [11], [38]. However, the empathy and transparency trade-off might not be mutually exclusive, as transparency can breed trust, which also influences interaction.

Limitations and Future Work

The primary limitation in our study design is that not all participants were exposed to all conditions. Given the number of conditions (varying generation/retrieval and transparent/not transparent author), we opted to mix within-subject comparisons and cross-study comparisons, resulting in a less clean study design. However, given the size of our online study, with around 200 participants per study, our results are still statistically sound. Future work can aim to replicate our findings with different study designs in order to confirm soundness of the psychological insights.

In addition, given the nature of crowdsourcing and the demographic pool of participants we surveyed, it is important to ensure that findings are replicated in other diverse populations. While our studies were roughly balanced by gender, Prolific respondents are predominantly white. Future work can assess the impact of identity on empathetic reaction to stories told by AI systems.

Conclusions

A growing number of companies and research institutions propose using language models and AI chatbots to improve mental wellbeing or social companionship. Empathy is a core tenet at the center of these chatbot designs, making it crucial to consider the ethical question of how empathy unfolds towards human vs AI narrators, and the role of transparency in this effect. To this end, we conducted four crowdsourced studies to assess how empathy differs across human-written vs AI-written stories, varying how stories are selected (generation vs retrieval) and author disclosure (transparency that story was written by an AI author vs. no transparency). While we utilize current state-of-the-art empathetic retrieval and generation in this work, our findings provide more generalized future insights around human behavior when interacting with AI chatbots. We find that transparency of the author plays an important role in empathy towards an AI story as well as people’s willingness to empathize towards machines. Our work motivates future directions regarding the social, psychological, and ethical implications of nuanced AI system design considerations that can drastically affect the ways in which humans extend empathy to artificial agents in the broader mental health and social support domains.

Acknowledgements

We would like to thank all of our participants, annotators, and teammates for their invaluable contributions to this project. This work was supported by an NSF GRFP under Grant No. 2141064 and the IITP grant funded by the Korean Ministry of Science and ICT No.2020-0-00842.

Conflicts of Interest

None declared.

Abbreviations

LLM: large language model

References

[1] J. Decety and P. L. Jackson, “The Functional Architecture of Human Empathy,” *Behavioral and Cognitive Neuroscience Reviews*, vol. 3, no. 2, pp. 71–100, Jun. 2004, doi: 10.1177/1534582304267187.

[2] B. M. P. Cuff, S. J. Brown, L. Taylor, and D. J. Howat, “Empathy: A Review of the Concept,” *Emotion Review*, vol. 8, no. 2, pp. 144–153, Apr. 2016, doi: 10.1177/1754073914558466.

[3] E. Cho and S. Jeon, “The role of empathy and psychological need satisfaction in pharmacy students’ burnout and well-being,” *BMC Medical Education*, vol. 19, no. 1, p. 43, Feb. 2019, doi: 10.1186/s12909-019-1477-2.

[4] V. Ta *et al.*, “User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis,” *Journal of Medical Internet Research*, vol. 22, no. 3, p. e16235, Mar. 2020, doi: 10.2196/16235.

[5] B. Inkster, S. Sarda, and V. Subramanian, “An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study,” *JMIR mHealth and uHealth*, vol. 6, no. 11, p. e12106, Nov. 2018, doi: 10.2196/12106.

[6] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial,” *JMIR mental health*, vol. 4, no. 2, p. e19, Jun. 2017, doi: 10.2196/mental.7785.

[7] M. Spitale, S. Okamoto, M. Gupta, H. Xi, and M. J. Matarić, “Socially Assistive Robots as Storytellers That Elicit Empathy,” *ACM Transactions on Human-Robot Interaction*, p. 3538409, May 2022, doi: 10.1145/3538409.

[8] K. Schaaff, C. Reinig, and T. Schlippe, “Exploring ChatGPT’s Empathic Abilities.” arXiv, Sep. 2023. doi: 10.48550/arXiv.2308.03527.

[9] C. Pelau, D.-C. Dabija, and I. Ene, “What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry,” *Computers in Human Behavior*, vol. 122, p. 106855, Sep. 2021, doi: 10.1016/j.chb.2021.106855.

[10] Y. Lee, M. Ha, S. Kwon, Y. Shim, and J. Kim, “Egoistic and altruistic motivation: How to induce users’ willingness to help for imperfect AI,” *Computers in Human Behavior*, vol. 101, pp. 180–196, Dec. 2019, doi: 10.1016/j.chb.2019.06.009.

[11] X. Lv, Y. Yang, D. Qin, X. Cao, and H. Xu, “Artificial intelligence service recovery: The role of empathic response in hospitality customers’ continuous usage intention,” *Computers in Human Behavior*, vol. 126, p. 106993, Jan. 2022, doi: 10.1016/j.chb.2021.106993.

[12] L. L. Chung and J. Kang, “\I’m Hurt Too\: The Effect of a Chatbot\s Reciprocal Self-Disclosures on Users’ Painful Experiences,” *Archives of Design Research*, vol. 36, no. 4, pp. 67–84, Nov. 2023, doi: 10.15187/adr.2023.11.36.4.67.

[13] K. Oatley, “Fiction: Simulation of Social Worlds,” *Trends in Cognitive Sciences*, vol. 20, no. 8, pp. 618–628, Aug. 2016, doi: 10.1016/j.tics.2016.06.002.

[14] M. Djikic, K. Oatley, and M. C. Moldoveanu, "Reading other minds: Effects of literature on empathy," *Scientific Study of Literature*, vol. 3, no. 1, pp. 28–47, Jan. 2013, doi: 10.1075/ssol.3.1.06dji.

[15] G. Abercrombie, A. C. Curry, T. Dinkar, V. Rieser, and Z. Talat, "Mirages: On Anthropomorphism in Dialogue Systems." arXiv, Oct. 2023. doi: 10.48550/arXiv.2305.09800.

[16] M. Alonso, "Can Robots have Personal Identity?" *International Journal of Social Robotics*, Jan. 2023, doi: 10.1007/s12369-022-00958-y.

[17] S. Gabriel, I. Puri, X. Xu, M. Malgaroli, and M. Ghassemi, "Can AI Relate: Testing Large Language Model Response for Mental Health Support." arXiv, May 2024. Accessed: May 23, 2024. [Online]. Available: http://arxiv.org/abs/2405.12021

[18] P. Brandtzaeg, M. Skjuve, and A. Følstad, "My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship," *Human Communication Research*, vol. 48, Apr. 2022, doi: 10.1093/hcr/hqac008.

[19] E. Croes and M. Antheunis, "Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot," *Journal of Social and Personal Relationships*, vol. 38, p. 026540752095946, Sep. 2020, doi: 10.1177/0265407520959463.

[20] J. Zaki, *The war for kindness: Building empathy in a fractured world*. Crown, 2019. Accessed: Nov. 01, 2023. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=Bz11DwAAQBAJ&oi=fnd&pg=PA1&dq=info:TBdmyGO4bsIJ:scholar.google.com&ots=u7Qs2lRf8l&sig=5fccZ-HwyweuLDfgdLRCRcpMTic

[21] S. Giorgi, D. M. Markowitz, N. Soni, V. Varadarajan, S. Mangalik, and H. A. Schwartz, "'I Slept Like a Baby': Using Human Traits To Characterize Deceptive ChatGPT and Human Text," 2023.

[22] F. Ishowo-Oloko, J.-F. Bonnefon, Z. Soroye, J. Crandall, I. Rahwan, and T. Rahwan, "Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 517–521, Nov. 2019, doi: 10.1038/s42256-019-0113-5.

[23] C. L. V. Straten, J. Peter, R. Kühne, and A. Barco, "Transparency about a Robot's Lack of Human Psychological Capacities: Effects on Child-Robot Perception and Relationship Formation," *ACM Transactions on Human-Robot Interaction*, vol. 9, no. 2, pp. 1–22, Jun. 2020, doi: 10.1145/3365668.

[24] J. Shen, M. Sap, P. Colon-Hernandez, H. Park, and C. Breazeal, "Modeling empathic similarity in personal narratives," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6237–6252. doi: 10.18653/v1/2023.emnlp-main.383.

[25] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." arXiv, Aug. 2019. Accessed: Jun. 24, 2022. [Online]. Available: http://arxiv.org/abs/1908.10084

[26] R. Atkinson, "The Life Story Interview," p. 21, 1998.

[27] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT." Zenodo, 2020. doi:

10.5281/zenodo.4461265.

[28] L. Shen, “On a Scale of State Empathy During Message Processing,” *Western Journal of Communication*, vol. 74, no. 5, pp. 504–524, Oct. 2010, doi: 10.1080/10570314.2010.512278.

[29] M. Q. Patton, “Qualitative research,” *Encyclopedia of statistics in behavioral science*, 2005.

[30] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset,” arXiv, arXiv:1811.00207, Aug. 2019. doi: 10.48550/arXiv.1811.00207.

[31] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, “Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense.” arXiv, Oct. 2023. doi: 10.48550/arXiv.2303.13408.

[32] B. Liu, “In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction,” *Journal of Computer-Mediated Communication*, vol. 26, no. 6, pp. 384–402, Nov. 2021, doi: 10.1093/jcmc/zmab013.

[33] C. Jones and B. Bergen, “Does GPT-4 pass the turing test?” *arXiv preprint arXiv:2310.20216*, 2023.

[34] K. Schaaff, C. Reinig, and T. Schlippe, “Exploring chatgpt’s empathic abilities,” in *2023 11th international conference on affective computing and intelligent interaction (ACII)*, IEEE, 2023, pp. 1–8. doi: 10.48550/arXiv.2308.03527.

[35] C. Montemayor, J. Halpern, and A. Fairweather, “In principle obstacles for empathic AI: Why we can’t replace human empathy in healthcare,” *AI & SOCIETY*, vol. 37, no. 4, pp. 1353–1359, Dec. 2022, doi: 10.1007/s00146-021-01230-z.

[36] J. Hohenstein and M. Jung, “AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust,” *Computers in Human Behavior*, vol. 106, p. 106190, May 2020, doi: 10.1016/j.chb.2019.106190.

[37] R. V. Yampolskiy, “Taxonomy of Pathways to Dangerous AI.” arXiv, Nov. 2015. doi: 10.48550/arXiv.1511.03246.

[38] S. Jeong, L. Aymerich-Franch, S. Alghowinem, R. W. Picard, C. L. Breazeal, and H. W. Park, “A Robotic Companion for Psychological Well-being: A Long-term Investigation of Companionship and Therapeutic Alliance,” in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, Stockholm Sweden: ACM, Mar. 2023, pp. 485–494. doi: 10.1145/3568162.3578625.